



*Enteropathogen Resource Integration Center*  
Bioinformatics Resource Center

# **Natural Language Processing – Assisted Extraction of Information from Literature**



# Approach

- Utilize SRA's natural language processing tool (NetOwl<sup>®</sup> Extractor) on the ERIC-BRC literature space to speed extraction of meaningful information
  - SRA working on NLP since 1986, NetOwl first released by SRA in 1996.
  - Used by many fields including financial, intelligence.
  - Robust and well accepted (National Institute of Standards and Technology "Automated Content Extractions (ACE)" trials, top scorer for the past three years)
- Application for ERIC:
  - Create training and blind sets
  - Write patterns to extract information
  - Build interface so scientists can use information to assist annotation



# Information to Extract - entities

- Genes and gene products: (invC and InvC)
- Operons: (paaABCDE, inv locus)
- Organism: (*E. coli*, *Shigella flexneri*, human)
- Strain: (K-12, LT2)
- Non-enzyme protein: (prepilin, actin, gluconate transporter)
- Enzymes: (CDP-tyvelose epimerase, cysteine desulfurase)



# Information to Extract - links

- Relationships between entities and what the entities do.
- Three types
  1. (gene, gene product, operon) to role
  2. (gene, gene product, operon) mutation to result
  3. organism to pathogenesis
- Extraction based on syntactical structure

# Gene to Role

- “These results suggest that VirK function is an essential virulence determinant for shigellae involved in the expression of virG gene product at post-transcriptional level.”
  - VirK - is an essential ... post-transcriptional level
- “The sitA gene encodes a putative periplasmic binding protein, sitB encodes an ATP-binding protein, and sitC and sitD encode two putative permeases (integral membrane proteins).”
  - sitA - encodes a putative periplasmic binding protein
  - sitB - encodes an ATP-binding protein
  - sitC & sitD - encodes two putative permeases (integral membrane proteins)
- “including catalase-peroxidase (KatY), murine toxin (Ymt).”
  - KatY - catalase-peroxidase
  - Ymt - murine toxin

# Mutation to Result

- “After inactivation of icsB, the mutant strain remained invasive, but formed abnormally small plaques on HeLa cell monolayers, colonized only the peripheral cells of Caco-2 islets, and was unable to provoke a keratoconjunctivitis in guinea-pigs.”
  - icsB - the mutant strain remained invasive
  - icsB - formed abnormally small plaques on HeLa cell monolayers
  - icsB - colonized only the peripheral cells of Caco-2 islets
  - icsB - was unable to provoke a keratoconjunctivitis in guinea-pigs
- “The xdhA mutant grew faster with aspartate as a nitrogen source. The mutant also exhibited sensitivity to adenine, which guanosine partially reversed.”
  - xdhA – grew faster with aspartate as a nitrogen source.
  - xdhA – exhibited sensitivity to adenine, which guanosine partially reduced

# Organism to pathogenesis

- “Because S. typhimurium causes gastroenteritis in both cattle and humans, we believe that this information may be directly applicable to the human.”
  - S. typhimurium - causes gastroenteritis in both cattle and humans
- “Shigella flexneri causes bacillary dysentery by invading epithelial cells of the colonic mucosa”
  - Shigella flexneri - causes bacillary dysentery



# A Proposed Use

**Enteropathogen Resource Integration Center**  
Bioinformatics Resource Center

[Home](#) [About ERIC](#) [News](#) [Links](#) [ASAP](#) [Enteropathogens](#) [Publications](#)

Diarrheagenic E. coli Shigella Salmonella Yersinia enterocolitica Yersinia pestis

**PMID: 10452520**

[Back to Today's PubMeds](#)

S: FEBS Lett. 1999 Jul 30;456(1):13-6.

Knock-out of the **cyaY** gene in **Escherichia coli** does not affect cellular iron content and sensitivity to oxidants.

Li DS, Ohshima K, Jiralerspong S, Bojanowski MW, Pandolfo M.

Research Center, Centre Hospitalier de l'Universite de Montreal, Hopital Notre-Dame, Montreal, Que., Canada.

Friedreich ataxia is a recessively inherited neurodegenerative disease caused by deficiency of a highly conserved mitochondrial protein, **frataxin**. **Frataxin** deficiency results in mitochondrial iron accumulation and oxidative stress. **Frataxin** shows homology with the **CyaY** proteins of **gamma-purple bacteria**, whose function is unknown. We knocked out the **CyaY** gene in **Escherichia coli** **MM383** by homologous recombination and we generated an **E. coli** **MM383** strain overexpressing **CyaY**. Bacterial growth, iron content and survival after exposure to H2O2 did not differ among these strains, suggesting that, despite structural similarities, **cyaY** proteins in bacteria may have a different function from **frataxin** homologues in mitochondria.

PMID: [10452520](#) [PubMed - indexed for MEDLINE]

**Highlighted abstract**

**Gene Roles determined by PMID:10452520**

**cyaY**

- may have a different function from **frataxin** homologues in mitochondria [\[Annotate\]](#)

**Gene Products determined by PMID:10452520**

**frataxin**

- a highly conserved mitochondrial protein

**Mutation Results determined by PMID:10452520**

**cyaY**

- does not affect cellular iron content and sensitivity to oxidants

**Frataxin**

- results in mitochondrial iron accumulation and oxidative stress

**Summary for PMID:10452520**

**Organism**

Name	Occurs	Color?
<a href="#">E. coli</a>	3	<input checked="" type="checkbox"/>
<a href="#">gamma-purple bacteria</a>	1	<input checked="" type="checkbox"/>

**Strain**

Name	Occurs	Color?
<a href="#">MM383</a>	2	<input checked="" type="checkbox"/>

**Enzyme**

Name	Occurs	Color?
- None -	-	-

**Gene**

Name	Occurs	Color?
<a href="#">cyaY</a>	5	<input checked="" type="checkbox"/>

**Operon**

Name	Occurs	Color?
- None -	-	-

**Nucleic Acid**

Name	Occurs	Color?
- None -	-	-

**Non Enzyme Protein**

Name	Occurs	Color?
<a href="#">frataxin</a>	4	<input checked="" type="checkbox"/>

**Extracted data**

**Summary counts**



# A Proposed Use - detail

**Gene Roles determined by PMID:10048039**

**sirA**

- is a global regulator of genes mediating enteropathogenesis [\[Annotate\]](#)
- is known to regulate the **hilA** and **prgH** genes within **Salmonella** pathogenicity island 1 (SPI1) [\[Annotate\]](#)
- positively regulated fusions [\[Annotate\]](#)
- regulates **hilA** [\[Annotate\]](#)
- regulates genes within SPI1 [\[Annotate\]](#)

**hilA**

- an SPI1 transcriptional regulator [\[Annotate\]](#)
- controlling enteropathogenic virulence functions in **S. typhimurium** [\[Annotate\]](#)

**spaS**

- a component of the SPI1 type III export apparatus [\[Annotate\]](#)

**sipB**

- a substrate of the SPI1 export apparatus [\[Annotate\]](#)

**sopB**

- is located within SPI5 [\[Annotate\]](#)
- is exported via the SPI1 export apparatus [\[Annotate\]](#)

**Gene Products determined by PMID:10048039**

- None Identified -

**Mutation Results determined by PMID:10048039**

**sirA**

- are dramatically attenuated in a **bovine** model of gastroenteritis
- have little or no effect in the **mouse** model of typhoid fever

**hilA**

- are dramatically attenuated in a **bovine** model of gastroenteritis
- have little or no effect in the **mouse** model of typhoid fever

typhimurium 3 ☒

**Enzyme**

Name	Occurs	Color?
- None -	-	-

**Gene**

Name	Occurs	Color?
sirA	8	<input checked="" type="checkbox"/>
hilA	9	<input checked="" type="checkbox"/>
prgH	1	<input checked="" type="checkbox"/>
mudJ	1	<input checked="" type="checkbox"/>
spaS	2	<input checked="" type="checkbox"/>
sopB	2	<input checked="" type="checkbox"/>
siqD	1	<input checked="" type="checkbox"/>
sipB	2	<input checked="" type="checkbox"/>

**Operon**

Name	Occurs	Color?
lacZY	1	<input checked="" type="checkbox"/>

**Nucleic Acid**

Name	Occurs	Color?
- None -	-	-

**Non Enzyme Protein**

Name	Occurs	Color?
- None -	-	-

Display settings

Links to existing annotations



# Example Workflow – Examining Current Literature

My PubMeds

Today Search

Organism/Gene Name:

Results:

- PMID 10048039: *Salmonella* SirA is a global regulator of genes mediating enteropathogenesis ... regulate the *hila* and *prgH* genes within *Salmonella* pathogenicity island 1 (SPI1). ... the *hila* (...), *spa5* (...) and *sipB* (... *sopB* gene (also known as *sigD*) ... *sopB* is located w dependent, ... (except *hila*) for *hila* dependence. ... require *hila* for expression and that p SirA regulates *hila*, ... Both *sirA* and *hila* mutants are dramatically ... the *SirA/Hila* regu typhimurium causes gastroenteritis in both cattle and humans...
- PMID 10085045: *Salmonella typhimurium* encodes a putative iron transport system within th Upon entry into the host, *Salmonella enterica* strains are presumed to ... pathogenicity is typhimurium chromosome. This locus, designated *sit*, ... transport system of *Yersinia pe* and *sitC* and *sitD* encode two ... defect of the enterobactin-deficient *Escherichia coli* strai operon is ... Introduction of a *sitBCD* deletion into wild-type *S. typhimurium* resulted in no

Enteropathogen Resource Integration Center  
Bioinformatics Resource Center

Home About ERIC News Links ASAP Enteropathogens Publications

Diarrheagenic E. coli Shigella Salmonella Yersinia enterocolitica Yersinia pestis

PMID: 10452520

Back to Today's PubMeds

S: FEBIS LAM, 1999 Jul 20;45(11):13-9.  
Knock-out of the *cydA* gene in *Escherichia coli* does not affect cellular iron content and sensitivity to oxidants.  
Li DS, Ohshima K, Jindalasing S, Bijzeman JH, Pandolfi M.  
Research Center, Centre Hospitalier de l'Université de Montréal, Hôpital Notre-Dame, Montréal, Que., Canada.  
Friedreich ataxia is a recessively inherited neurodegenerative disease caused by deficiency of a highly conserved mitochondrial protein, *frataxin*. *Frataxin* deficiency results in mitochondrial iron accumulation and oxidative stress. *Frataxin* shows homology with the *cydA* proteins of *gamma-proteobacteria*, whose function is bacterial iron content and survival after exposure to H<sub>2</sub>O<sub>2</sub> did not differ among these strains, suggesting that, despite structural similarities, *cydA* proteins in bacteria may have a different function from *frataxin* homologues in mitochondria.  
PMID: 10452520 [PubMed - in process] MEDLINE]

Gene Roles determined by PMID: 10452520

**cydA**  
• may have a different function from *frataxin* homologues in mitochondria. [Annotate]

Gene Products determined by PMID: 10452520

**Frataxin**  
• a highly conserved mitochondrial protein.

Mutation phenotypes determined by PMID: 10452520

**cydA**  
• does not affect cellular iron content and sensitivity to oxidants

**Frataxin**  
• results in mitochondrial iron accumulation and oxidative stress

Summary for PMID: 10452520

Organism

Name Occurs Color?

*E. coli* 1 [X]

*gamma-proteobacteria* 1 [X]

Strain

Name Occurs Color?

PM103 2 [X]

Enzyme

Name Occurs Color?

None - [X]

Gene

Name Occurs Color?

*cydA* 1 [X]

Operon

Name Occurs Color?

None - [X]

Nucleic Acid

Name Occurs Color?

None - [X]

Non Enzyme Protein

Name Occurs Color?

Frataxin 4 [X]

- Today's relevant abstracts processed through system. Interface will present these results to scientific community allowing them to drill down to detail
- Interface links to existing annotations allowing user to determine what is new in this article, add and modify annotations
- Can be either all articles or from a set of keywords (eg. specific genes)

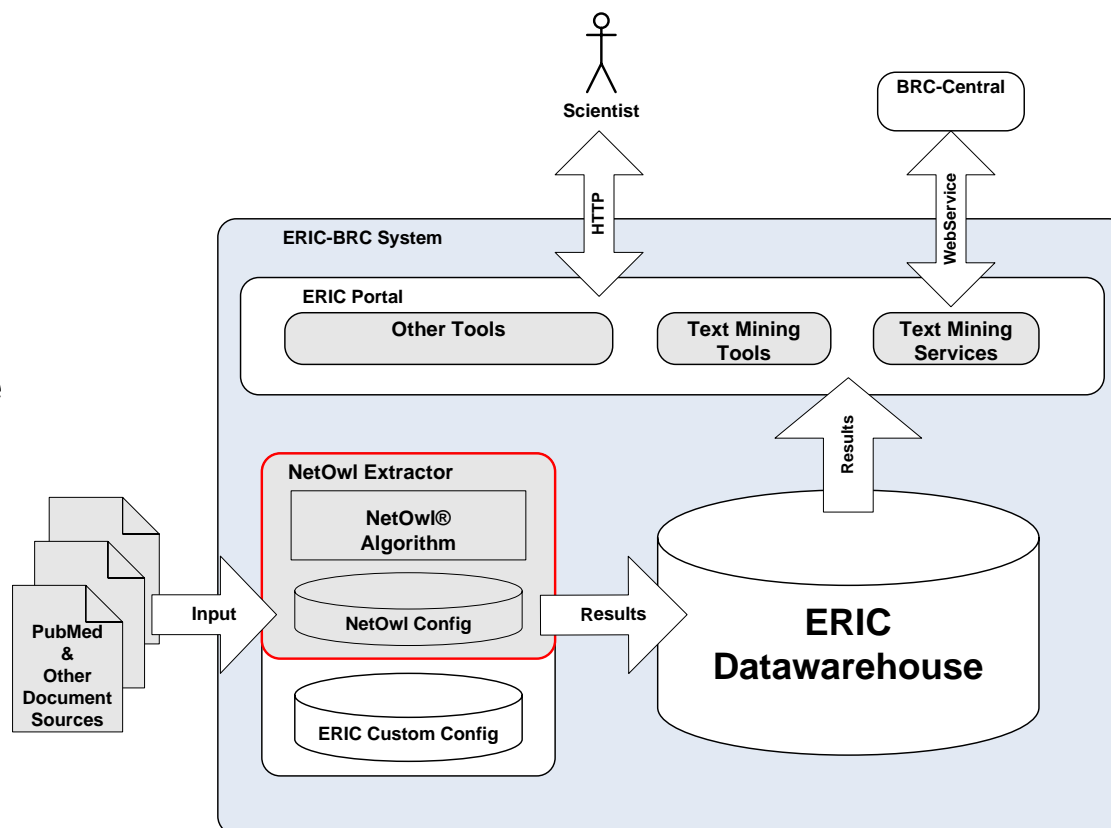


# NetOwl Extractor

## Optimizing Manual Literature Annotation



- NetOwl is privately developed software by SRA International.
- Being configured to mine information from abstracts and journals articles relevant to ERIC.
- Will increase efficiency by identifying interesting information for annotators.



\*Software licensed for use on ERIC bounded in red.